RNA-Seq workflow

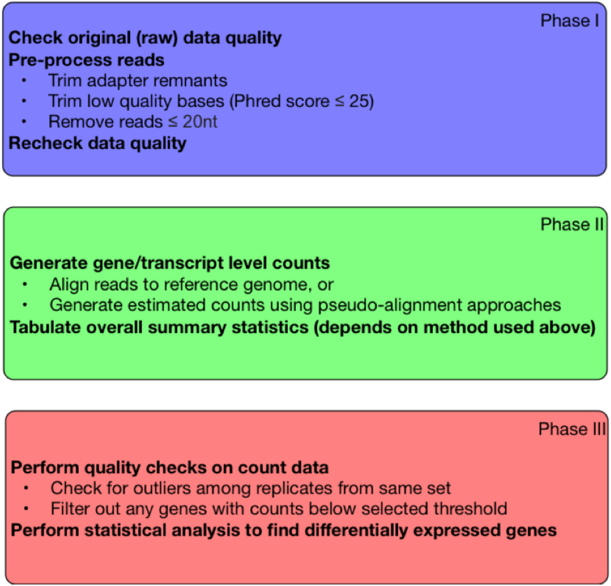
Phase 2: Generating expression count data

Andrew Ndhlovu

Project SeaStore
Department of Botany and Zoology
Stellenbosch University
https://github.com/PiscatorX/RNA-Seq-devs

July 24, 2023

Overview of RNA-Seq workflow



Phase I

Check original (raw) data quality

Pre-process reads

- Trim adapter remnants
- Trim low quality bases (Phred score ≤ 25)
- Remove reads ≤ 20nt

Recheck data quality

Phase II

Generate gene/transcript level counts

- Align reads to reference genome, or
- Generate estimated counts using pseudo-alignment approaches

Tabulate overall summary statistics (depends on method used above)

Phase III

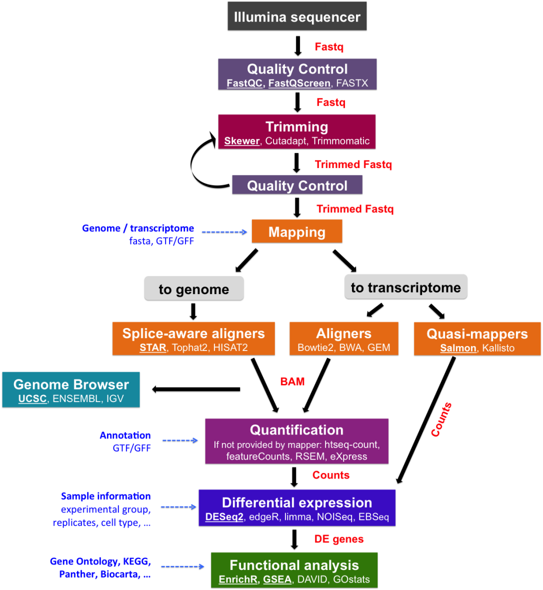
Perform quality checks on count data

- Check for outliers among replicates from same set
- Filter out any genes with counts below selected threshold

Perform statistical analysis to find differentially expressed genes

https://h3abionet.github.io/H3ABionet-SOPs/RNA-Seq-1-2.html

RNA-Seq goal is differential expression analysis



Illumina sequencer

Fastq

Quality Control
FastQC, FastQScreen, FASTX

Fastq

Trimming
Skewer, Cutadapt, Trimmomatic

Trimmed Fastq

Quality Control

Trimmed Fastq

Genome / transcriptome
fasta, GTF/GFF

Mapping

to genome

to transcriptome

Splice-aware aligners
STAR, Tophat2, HISAT2

Aligners
Bowtie2, BWA, GEM

Quasi-mappers
Salmon, Kallisto

Genome Browser
UCSC, ENSEMBL, IGV

BAM

Counts

Annotation
GTF/GFF

Quantification
If not provided by mapper: htseq-count, featureCounts, RSEM, eXpress

Counts

Sample information
experimental group, replicates, cell type, …

Differential expression
DESeq2, edgeR, limma, NOISeq, EBSeq

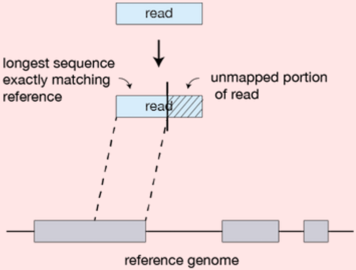
DE genes

Gene Ontology, KEGG, Panther, Biocarta, …

Functional analysis
EnrichR, GSEA, DAVID, GOstats

RNA-Seq alignment is key to expression quantification

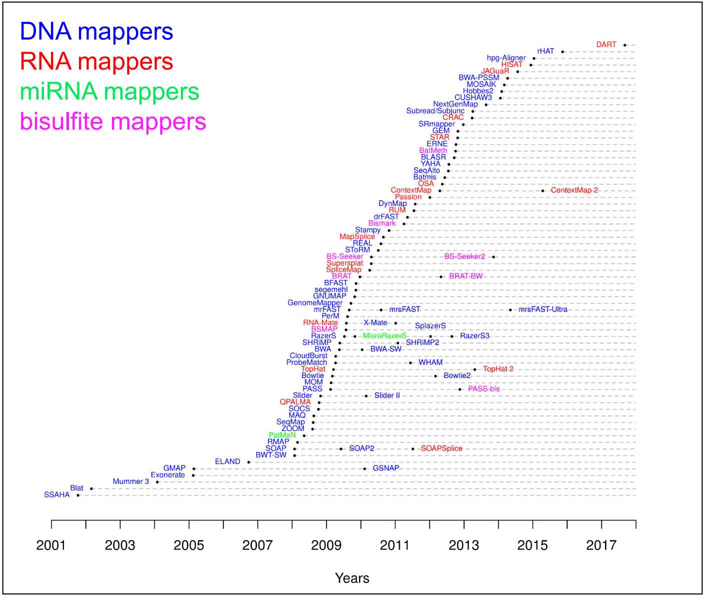
Challenges of aligning RNA-seq reads to a genome



https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

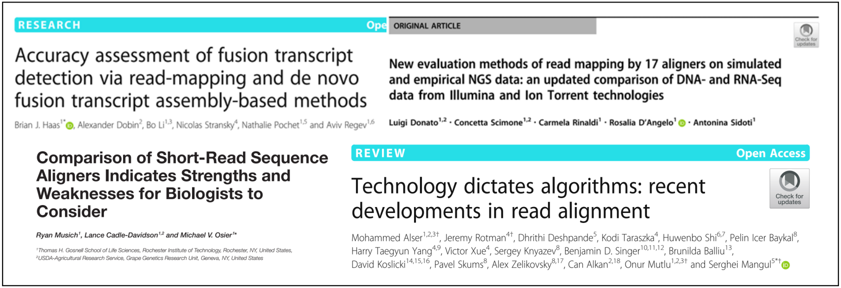
- RNA reads do not contain introns (spliced out during transcription)
- Sequences may span multiple exons
- Mapping to a reference genome is computationally challenging
- Too many reads: millions to align!!!
- Need for splice aware aligners

Spoilt for choice: Lots of mappers



http://genomeast.igbmc.fr/wiki/lib/exe/fetch.php?media=training:190115_rnaseq_phdpgr.pdf

Lots of comparison studies



RESEARCH Ope

Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods

Brian J. Haas[1*], Alexander Dobin[2], Bo Li[1,3], Nicolas Stransky[4], Nathalie Pochet[1,5] and Aviv Regev[1,6]

ORIGINAL ARTICLE

Check for updates

New evaluation methods of read mapping by 17 aligners on simulated and empirical NGS data: an updated comparison of DNA- and RNA-Seq data from Illumina and Ion Torrent technologies

Luigi Donato[1,2] · Concetta Scimone[1,2] · Carmela Rinaldi[1] · Rosalia D'Angelo[1] · Antonina Sidoti[1]

Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider

Ryan Musich[1], Lance Cadle-Davidson[1,2] and Michael V. Osier[1]*

[1] Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester, NY, United States,
[2] USDA-Agricultural Research Service, Grape Genetics Research Unit, Geneva, NY, United States

REVIEW Open Access

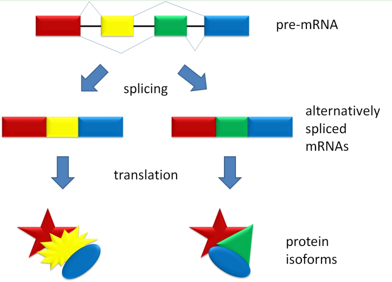
Technology dictates algorithms: recent developments in read alignment

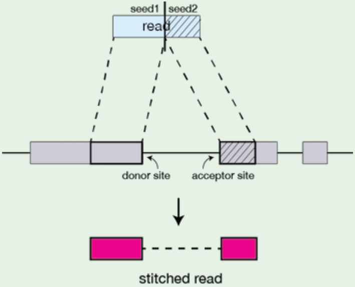
Check for updates

Mohammed Alser[1,2,3†], Jeremy Rotman[4†], Dhrithi Deshpande[5], Kodi Taraszka[4], Huwenbo Shi[6,7], Pelin Icer Baykal[8], Harry Taegyun Yang[4,9], Victor Xue[4], Sergey Knyazev[8], Benjamin D. Singer[10,11,12], Brunilda Balliu[13], David Koslicki[14,15,16], Pavel Skums[8], Alex Zelikovsky[8,17], Can Alkan[2,18], Onur Mutlu[1,2,3†] and Serghei Mangul[5*†]

STAR: Spliced Transcripts Alignment to a Reference

A fast splice aware aligner



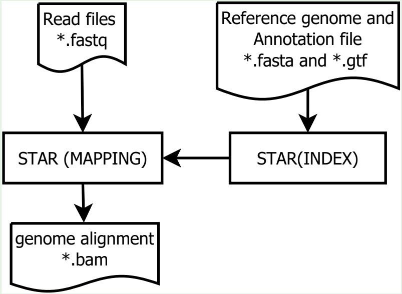


https://simple.wikipedia.org/wiki/Alternative_splicing https://www.reneshbedre.com/blog/star-aligner.html

- STAR outperforms other aligners by $>$ 50X in mapping speed
- Discovers non-canonical splices and chimeric (fusion) transcripts
  - Aligns reads with indels due to genomic variations or sequencing errors.
  - Identifies spliced RNAs formed by sequences across genomic regions

https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf

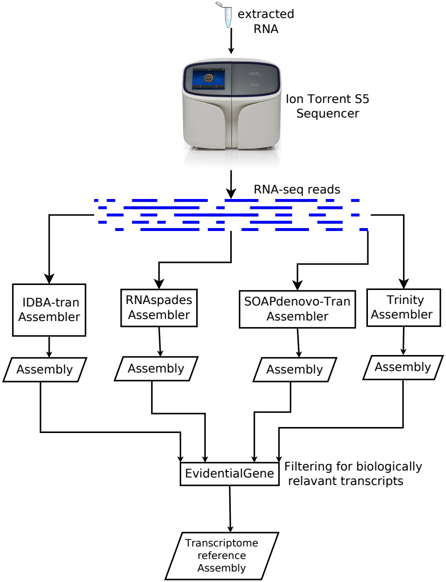
Basic STAR workflow consists of 2 steps



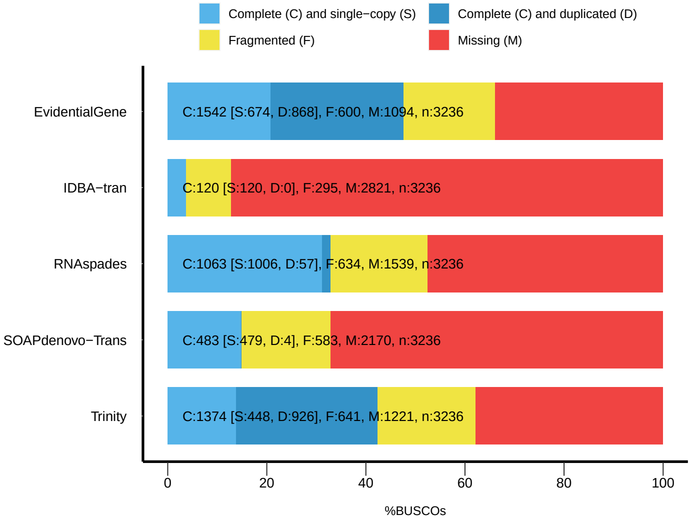
1. Generating genome indexes files.
   Reference genome sequences (FASTA files) and annotations (GTF file) to generate genome indexes
2. Mapping reads to the genome.
   STAR maps the reads to the genome index, and writes several output files, such as alignments (SAM/BAM), mapping summary statistics, splice junctions, unmapped reads, signal (wiggle) tracks etc.

https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf

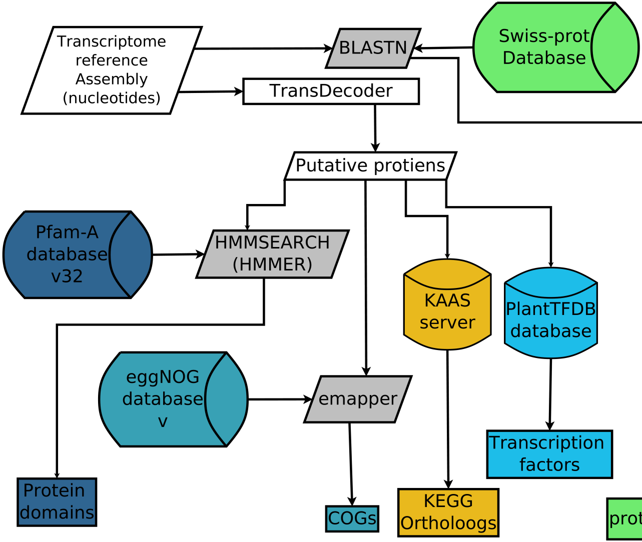
Sequencing and *denovo* transcriptome assembly



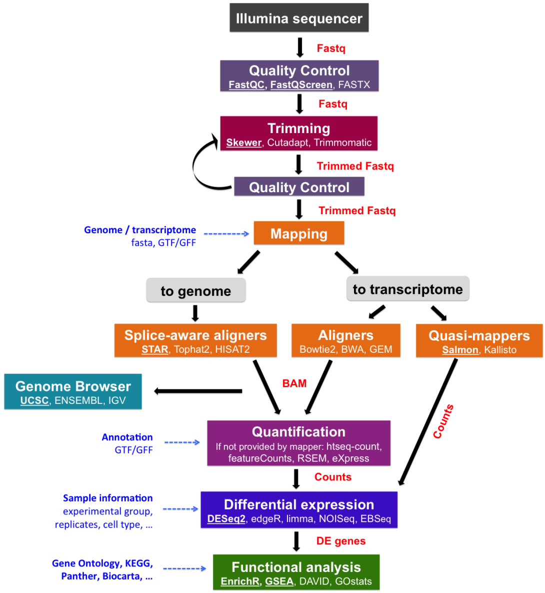
Benchmarking Universal Single-Copy Orthologs (BUSCOs)



Annotation of transcripts and predicted proteins

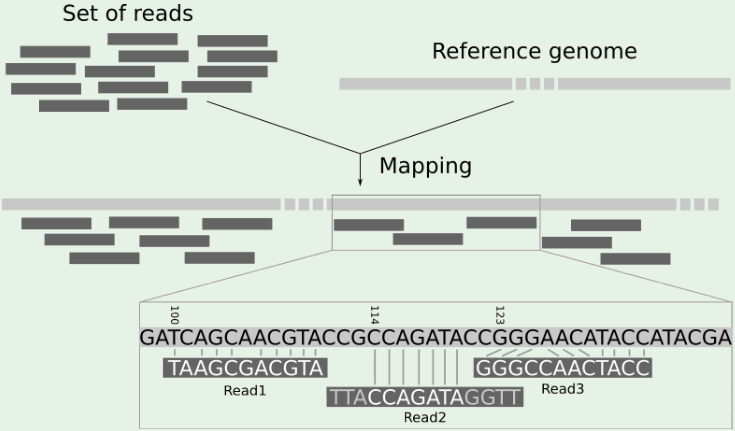


Recap: Goal of RNA-Seq is differential expression analysis



Mapping reads to the transcriptome reference

Bowtie 2: an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences



`https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html`

Read mapping workflow

Bowtie

`https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml`

- Building an index from the transcriptome reference
- Aligning reads to reference. Output is a sequence alignment/map (SAM) file

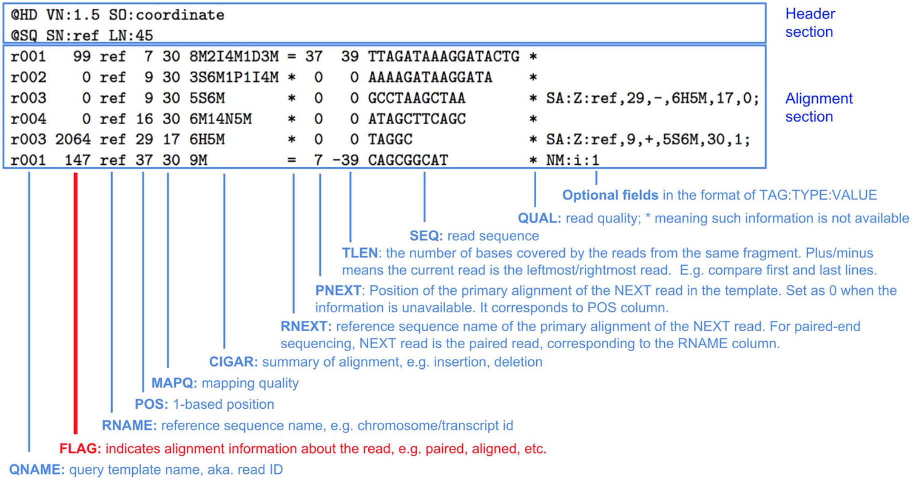
samtools

`http://www.htslib.org/`

- Convert SAM to BAM (binary alignment map) file
  [`samtools view -b <sam>`]
- `Sort BAM files using referenc [ samtools sort <bam> ]`
- `Index BAM [ samtools index <bam> ]`
- `Get alignment statistics [ samtools flagstat <bam> ]`

What is in a SAM file

https://samtools.github.io/hts-specs/SAMv1.pdf



```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Header section

Alignment section

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; * meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

POS: 1-based position

RNAME: reference sequence name, e.g. chromosome/transcript id

FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID

https://www.samformat.info/sam-format-flag

Inspecting the bowtie mapping

Check samtools stats

`https://davetang.org/wiki/tiki-index.php?page=SAMTools`

```
samtools flagstat <bam>
```

Visualiase mapping using IGV

`https://software.broadinstitute.org/software/igv/` Integrative Genomics Viewer (IGV) is an interactive tool for the visual exploration of genomic data.

- bam file
- bam file index
- reference sequence